

Some Insights About the Small Ball Probability Factorization for Hilbert Random Elements

Enea G. Bongiorno, Aldo Goia
 Università del Piemonte Orientale,
 enea.bongiorno@uniupo.it, aldo.goia@uniupo.it

March 30, 2016

Abstract

Asymptotic factorizations for the small-ball probability (SmBP) of a Hilbert valued random element X are rigorously established and discussed. In particular, given the first d principal components (PCs) and as the radius ε of the ball tends to zero, the SmBP is asymptotically proportional to (a) the joint density of the first d PCs, (b) the volume of the d -dimensional ball with radius ε , and (c) a correction factor weighting the use of a truncated version of the process expansion. Moreover, under suitable assumptions on the spectrum of the covariance operator of X and as d diverges to infinity when ε vanishes, some simplifications occur. In particular, the SmBP factorizes asymptotically as the product of the joint density of the first d PCs and a pure volume parameter. All the provided factorizations allow to define a *surrogate intensity* of the SmBP that, in some cases, leads to a genuine *intensity*. To operationalize the stated results, a non-parametric estimator for the *surrogate intensity* is introduced and it is proved that the use of estimated PCs, instead of the true ones, does not affect the rate of convergence. Finally, as an illustration, simulations in controlled frameworks are provided.

Keywords. Hilbert functional data; Small Ball Probability; Karhunen–Loève decomposition; kernel density estimate.

Introduction

For a random element X valued in a general metric space, the measure of how it concentrates over such space plays a central role in statistical analysis. If X is a real random vector, its joint density is, in a natural way, that measure. In fact, in practical situations, the density is helpful in defining mixture models, in detecting latent structure, in discriminant analysis, in robust statistics to identify outliers and so on. When observed data are curves, surfaces, images, objects or, briefly, *functional data* (see e.g. monographs Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Ramsay and Silverman, 2005, and Bongiorno et al., 2014 for recent contributions), the dimensionality of the

space to which the data belong raises problems in defining an object that plays the same role of the joint density distribution in the multivariate context. The main problem is that, without an underlying dominant probability measure, the Radon–Nikodym derivative can not be straightforwardly applied. To manage this, a concept of “surrogate density” is derived from the notion of small–ball probability (SmBP in the sequel) of a random element X , that is

$$\varphi(x, \varepsilon) = \mathbb{P}(\Delta(X, x) < \varepsilon),$$

where x is in the same space where X takes its values, ε is a positive real number and Δ is a semi–metric. The behaviour, as ε vanishes, of $\varphi(x, \varepsilon)$ provides information about the way in which X concentrates at x . Such feature stimulates the study of the SmBP in different settings: the limiting behaviour has been developed from a theoretical point of view (for instance, refer to the small tails/deviations theory Li and Shao, 2001; Lifshits, 2012 and references therein), in functional statistics the SmBP was used to derive asymptotics in mode estimations (see, e.g. Dabo-Niang et al., 2007; Delaigle and Hall, 2010; Ferraty et al., 2012; Gasser et al., 1998), as well as in non–parametric regression literature in evaluating the rate of convergence of estimators (see, e.g. Ferraty and Vieu, 2006; Ferraty et al., 2007).

Often, the necessity to have available a surrogate density for X has brought to assume (as done, for instance, in Ferraty et al., 2012; Gasser et al., 1998) that

$$\varphi(x, \varepsilon) = \Psi(x) \phi(\varepsilon) + o(\phi(\varepsilon)), \quad \varepsilon \rightarrow 0, \quad (1)$$

where Ψ , depending only on the center x , plays the role of the *surrogate density* of the random element X , whilst $\phi(\varepsilon)$ is a kind of “volume parameter” which does not depend on the spatial term. It is worth noticing that Ψ is also the *intensity* of the SmBP.

Although to break the dependence on x and ε supplies a clear modelling advantage and the existence of $\Psi(x)$ is desirable (especially from a statistical perspective), factorization (1) can be derived only in particular settings. Notable examples are the case of Gaussian processes (e.g. Li and Shao, 2001; Lifshits, 2012 and references therein) and the one of fractal processes for suitable semi–norms Δ (e.g. Ferraty and Vieu, 2006, Chapter 13). Hence, a crucial task is to study some asymptotic factorizations of the SmBP that lead to a definition of its *intensity* or, at least, when it is not possible to completely isolate the dependence on x and ε , a *surrogate intensity*.

In what follows, we assume X to be a random element in an Hilbert space with Δ being the metric induced by the Hilbert norm and, without loss of generality, we deal with random curves on the space of square integrable functions on the unit interval. A first factorization of the SmBP that allows to define a *surrogate intensity* was provided by Delaigle and Hall (2010): besides some technical hypothesis, on the spectrum of the covariance operator of X and assuming that principal components of X are independent with positive and sufficiently smooth marginal density functions $\{\tilde{f}_j\}$, the authors showed that

$$\varphi(x, \varepsilon) \sim \prod_{j \leq d} \tilde{f}_j(x_j) \phi(\varepsilon, d), \quad \varepsilon \rightarrow 0, \quad (2)$$

where x_j is the projection of x over the j -th principal axis, $\phi(\varepsilon, d)$ is volumetric term and $d = d(\varepsilon)$ is the number of considered terms of the decomposition diverging to infinity as ε vanishes. Now, from the application point of view, the independence assumption appears quite restrictive and the spatial factor $\prod_{j \leq d} \tilde{f}_j$ results just a *surrogate intensity* of the SmBP because of the dependence between d and ε . Moreover, one may wonder if the principal component analysis is necessary to obtain the factorization.

The first part of the present work is devoted to propose more general factorizations for the SmBP that relax the hypothesis of independence, and to identify situations in which it is possible to obtain a genuine *intensity*. The first stated factorization holds for any positive integer d :

$$\varphi(x, \varepsilon) \sim f_d(x_1, \dots, x_d) V_d(\varepsilon) \mathcal{R}(x, \varepsilon, d), \quad \text{as } \varepsilon \rightarrow 0,$$

where f_d is the joint distribution of the first d principal components, $V_d(\varepsilon)$ is the volume of a d -dimensional ball with radius ε and $\mathcal{R}(x, \varepsilon, d) \in (0, 1]$ denotes an extra factor compensating the use of (x_1, \dots, x_d) instead of x . Such factorization benefits from the fact that d is fixed but, in general, a genuine *intensity* can not be defined because \mathcal{R} depends on both x and ε . Such dependency is bypassed and hence an *intensity* obtained if one introduces suitable assumptions on the probability law of the process and/or on the point x at which the factorization is evaluated.

Moving from this first factorization, we prove that

$$\varphi(x, \varepsilon) \sim f_d(x_1, \dots, x_d) \phi(\varepsilon, d), \quad \text{as } \varepsilon \rightarrow 0, \text{ and } d(\varepsilon) \rightarrow \infty,$$

where $\phi(\varepsilon, d)$ is a volume parameter that depends on the decay rate of $\{\lambda_j\}$, the eigenvalues of the covariance operator of X . Such result candidates, in a very natural way, the joint density distribution f_d to be a *surrogate intensity* of the SmBP and, under suitable assumptions on X , allows to define an *intensity*. Furthermore, it turns out that, the first factorization is basis free, while for the second one the principal components basis is optimal in some sense.

In the second part of the paper, in order to make available the *surrogate intensity* of the SmBP for statistical purpose, we propose a multivariate kernel density approach to estimate f_d . Under general conditions, we prove that, although the estimation procedure involves the estimated principal components instead of the true ones, the estimator achieves the classical non-parametric rate of convergence. To show how such estimator performs on finite sample frameworks, we study its behaviour by means of simulated processes whose *intensity* is known.

The paper outline goes as follow: Section 1 introduces the framework, Section 2 considers the factorization of the SmBP when d is fixed whereas Section 3 when d diverges to infinity as ε vanishes. Section 4 provides the statistical asymptotic theorem in estimating the joint density f_d . Section 5 illustrates some numerical examples. Finally Section 6 collects all the proofs.

1 Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{L}_{[0,1]}^2$ be the Hilbert space of square integrable real functions on $[0, 1]$ endowed with the inner product $\langle g, h \rangle = \int_0^1 g(t) h(t) dt$ and the

induced norm $\|g\|^2 = \langle g, g \rangle$. Consider a measurable map X defined on (Ω, \mathcal{F}) taking values in $(\mathcal{L}_{[0,1]}^2, \mathcal{B})$, where \mathcal{B} denotes the Borel sigma-algebra induced by $\|\cdot\|$. Denote by

$$\mu_X = \{\mathbb{E}[X(t)], t \in [0, 1]\}, \quad \text{and} \quad \Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)],$$

the mean function and covariance operator of X respectively. Let us consider the Karhunen–Loève expansion associated to X (see e.g. Bosq, 2000): denoting by $\{\lambda_j, \xi_j\}_{j=1}^\infty$ the decreasing to zero sequence of non-negative eigenvalues and their associated orthonormal eigenfunctions of the covariance operator Σ , the random curve X admits the following representation

$$X(t) = \mu_X(t) + \sum_{j=1}^{\infty} \theta_j \xi_j(t), \quad 0 \leq t \leq 1,$$

where $\theta_j = \langle X - \mu_X, \xi_j \rangle$ are the so-called principal components (PCs in the sequel) of X satisfying

$$\mathbb{E}[\theta_j] = 0, \quad \text{Var}(\theta_j) = \lambda_j, \quad \mathbb{E}[\theta_j \theta_{j'}] = 0, \quad j \neq j'.$$

It is just the case to recall that $\{\xi_j\}_{j=1}^\infty$ provides an orthonormal basis of the considered Hilbert space and that the Karhunen–Loève expansion, taking advantage of the euclidean underlying structure, isolates the manner in which the random function $X(\omega, t)$ depends upon t and upon ω .

In order to achieve our aims, let us consider the following assumptions.

(A-1) The process is centered, that is $\mu_X = 0$.

(A-2) The center of the ball $x \in \mathcal{L}_{[0,1]}^2$ is sufficiently close to the process in its high-frequency part, that is

$$x_j^2 \leq C_1 \lambda_j, \quad \text{for any } j \geq 1 \quad (3)$$

where $x_j = \langle x, \xi_j \rangle$ for some positive constant C_1 .

The latter is not a restrictive condition since it holds whenever x belongs to the reproducing kernel Hilbert space generated by the process X :

$$RKHS(X) = \left\{ x \in \mathcal{L}_{[0,1]}^2 : \sum_{j \geq 1} \lambda_j^{-1} \langle x, \xi_j \rangle^2 < \infty \right\}. \quad (4)$$

Roughly speaking, x is an element of $RKHS(X)$ only if it is “at least smooth as the covariance function”, see Berlinet and Thomas-Agnan, 2004, p. 13 and p. 69. Note that $RKHS(X)$ is a very large subspace of H including the finite dimensional ones; in fact, if $x_j = 0$ for any $j \geq d$ and any $d \in \mathbb{N}$, then $x \in RKHS(X)$. Furthermore, (A-2) is not unusual since it is equivalent to $\sup_{j \geq 1} \mathbb{E}[(\theta_j - x_j)^2 / \lambda_j] < \infty$ that was used, for similar purpose by Delaigle and Hall, 2010, Condition (4.1).

(A-3) Denote by Π_d the projector onto the d -dimensional space spanned by $\{\xi_j\}_{j=1}^d$. The first d PCs, namely $\boldsymbol{\theta} = \Pi_d X = (\theta_1, \dots, \theta_d)'$, admit a joint strictly positive probability density, namely $\boldsymbol{\vartheta} \in \mathbb{R}^d \mapsto f_d(\boldsymbol{\vartheta})$. Moreover, f_d is twice differentiable at $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_d)' \in \mathbb{R}^d$ and there exists a positive constant C_2 (not depending on d) for which

$$\left| \frac{\partial^2 f_d}{\partial \vartheta_i \partial \vartheta_j}(\boldsymbol{\vartheta}) \right| \leq \frac{C_2}{\sqrt{\lambda_i \lambda_j}} f_d(x_1, \dots, x_d), \quad (5)$$

for any $d \in \mathbb{N}$, $i, j \in \{1, \dots, d\}$ and $\boldsymbol{\vartheta} \in D^x = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j \leq d} (\vartheta_j - x_j)^2 \leq \rho^2 \right\}$ for some $\rho \geq \varepsilon$.

From now on, with a slight abuse of notation and when it is clear from the context, $f_d(x)$ denotes $f_d(x_1, \dots, x_d)$.

To better appreciate the meaning of (5), note that it can be derived in an intuitive way considering $\mathbf{W}^x = (W_1, \dots, W_d)'$, the deterministic translation of the component-wise standardized version of the PCs defined by

$$W_j^x = \frac{1}{\sqrt{\lambda_j}} \langle X - x, \xi_j \rangle = \frac{\theta_j - \langle x, \xi_j \rangle}{\sqrt{\lambda_j}}.$$

In fact, (5) is equivalent in assuming the boundedness of the second derivative of the density probability function g_d^x of the random vector \mathbf{W}^x . Since the latter is a linear transformation of $\boldsymbol{\theta}$, condition (5) is equivalent to

$$\left| \frac{\partial^2 g_d^x}{\partial w_i \partial w_j}(\mathbf{w}) \right| \leq C_2 g_d^x(\mathbf{0}),$$

for any $d \in \mathbb{N}$, $i, j \in \{1, \dots, d\}$ and $\mathbf{w} \in D' = \left\{ \mathbf{w} \in \mathbb{R}^d : \sum_{j \leq d} w_j^2 \lambda_j \leq \rho^2 \right\}$ for some $\rho \geq \varepsilon$. It is worth noting that (A-3) is not restrictive: it includes, for instance, the case of Gaussian Hilbert valued processes.

2 Approximation results for a given d

To state the main result of this section, let us consider a finite positive integer d , a given point $x \in \mathcal{L}_{[0,1]}^2$ and define

$$S = S(x, \varepsilon, d) = \frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2, \quad \mathcal{R}(x, \varepsilon, d) = \mathbb{E} \left[(1 - S)^{d/2} \mathbb{I}_{\{S < 1\}} \right], \quad (6)$$

and

$$V_d(\varepsilon) = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2 + 1)},$$

that is the volume of the d -dimensional ball with radius ε .

Theorem 1 *Let X be a process as above, $\varphi(x, \varepsilon)$ be the small ball probabilities of X , assume (A-1), ..., (A-3) and define*

$$\varphi_d(x, \varepsilon) = f_d(x) V_d(\varepsilon) \mathcal{R}(x, \varepsilon, d), \quad \text{for } \varepsilon > 0. \quad (7)$$

Then

$$|\varphi(x, \varepsilon) - \varphi_d(x, \varepsilon)| \leq C_2 \frac{\varepsilon^2}{2\lambda_d} \varphi_d(x, \varepsilon), \quad \text{for } \varepsilon > 0 \quad (8)$$

that is

$$\varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x, \varepsilon, d), \quad \text{for } \varepsilon \rightarrow 0. \quad (9)$$

In other words, for a fixed positive integer d and as $\varepsilon \rightarrow 0$, above theorem states that the SmBP $\varphi(x, \varepsilon)$ behaves as $\varphi_d(x, \varepsilon)$ (the usual first order approximation of the SmBP in a d -dimensional space $f_d(x) V_d(\varepsilon)$) up to the scale factor $\mathcal{R}(x, \varepsilon, d)$. The latter, depending on x only through its high-frequency components $\{x_j\}_{j \geq d+1}$, can be interpreted as a corrective factor compensating the use of a truncated version of the process expansion. Note that changing d affects all the terms in the factorization but not the asymptotic (9).

Because of $\mathcal{R}(x, \varepsilon, d)$, the dependence on x and ε can not be isolated in (9) and hence an *intensity* of the SmBP is not, in general, available. On the other hand, there exist some specific situations in which a genuine *intensity* can be defined from above factorization. We have identified three of these, namely:

- a) $\mathcal{R}(x, \varepsilon, d)$ is independent on x ,
- b) there exists a finite positive integer d_0 such that, for any $d \geq d_0$, $\mathcal{R}(x, \varepsilon, d) = 1$,
- c) for any x , $\begin{cases} \mathcal{R}(x, \varepsilon, d) \rightarrow 1, \\ \varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon), \end{cases} \quad \varepsilon \rightarrow 0, \quad d(\varepsilon) \rightarrow \infty.$

In the following, we discuss points a) and b), whereas point c) is discussed in Section 3 since requires additional arguments. In the last discussion we spend few words about the consequences of choosing a basis different from the Karhunen–Loève one.

D.1 – $\mathcal{R}(x, \varepsilon, d)$ is independent on x . Consider, for instance, $x_j = 0$ for any $j \geq d_0 + 1$ (i.e. x belongs to the space spanned by $\{\xi_1, \dots, \xi_{d_0}\}$). Then applying Theorem 1, for any $d \geq d_0$, we have

$$\varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(\varepsilon, d), \quad \text{as } \varepsilon \rightarrow 0,$$

where $V_d(\varepsilon) \mathcal{R}(\varepsilon, d)$ represents a pure volumetric term while f_d is an *intensity* of the SmBP evaluated at x . Let us consider the remarkable Gaussian process for which

$$\varphi(x, \varepsilon) \sim \exp \left\{ -\frac{1}{2} \sum_{j \leq d} \frac{x_j^2}{\lambda_j} \right\} \cdot \frac{V_d(\varepsilon) \mathcal{R}(\varepsilon, d)}{\prod_{j \leq d} \sqrt{2\pi\lambda_j}} = \Psi_d(x) \cdot \mathcal{V}_d(\varepsilon), \quad \text{as } \varepsilon \rightarrow 0,$$

where, for any $d \geq d_0$, $\Psi_d(x) = \Psi_{d_0}(x) = \exp \left\{ -\sum_{j \leq d_0} x_j^2 / (2\lambda_j) \right\}$ is the *intensity* of the SmBP evaluated at x . If we further specialize the above case to the Wiener process on $[0, 1]$, we can show that $\Psi_{d_0}(x)$ is coherent with already known results (see, for instance, Li and Shao, 2001, Theorem 3.1 and Dereich et al., 2003, Example 5.1). In fact, the Karhunen–Loève decomposition of a Wiener process is known to be

$$W(t) = \sum_{j=1}^{\infty} Z_j \xi_j(t), \quad \text{with} \quad \begin{cases} t \in [0, 1], \\ Z_j \sim N(0, 1) \text{ i.i.d.}, \\ \xi_j(t) = \sqrt{2} \sin((j - 0.5)\pi t) / \sqrt{\lambda_j}, \\ \lambda_j = (j - 0.5)^{-2} \pi^{-2}, \end{cases}$$

and it is known that

$$\varphi(x, \varepsilon) \sim \exp \left\{ -\frac{1}{2} \int_0^1 x'(t)^2 dt \right\} \frac{4\varepsilon}{\sqrt{\pi} \exp \left\{ \frac{1}{8\varepsilon^2} \right\}}, \quad \varepsilon \rightarrow 0,$$

where $x(t)$ is sufficiently smooth. Thus, the latter must be equivalent to $\Psi_{d_0}(x) \cdot \mathcal{V}_d(\varepsilon)$ as ε goes to zero and for any $d \geq d_0$. Since we are interested in the definition of an intensity, we do not care about the volumetric part and we just compare the spatial parts. Given $x(t) = \sum_{j=1}^{d_0} b_j \xi_j(t)$ where $b_j \in \mathbb{R}$, then x_j equals b_j , for $j = 1, \dots, d_0$ and is null otherwise. Moreover, straightforward computations lead to

$$\exp \left\{ -\frac{1}{2} \int_0^1 x'(t)^2 dt \right\} = \exp \left\{ -\frac{1}{2} \sum_{j=1}^{d_0} b_j^2 \right\} = \Psi_{d_0}(x).$$

The special case of Wiener process is exploited in Section 5.2 to construct numerical examples in estimating the *intensity* $\Psi_{d_0}(x)$.

D.2 – The case $\mathcal{R}(x, \varepsilon, d) = 1$. Consider X a d_0 -dimensional process (that is when it takes values in a d_0 -dimensional subspace of the Hilbert space). Then $\lambda_j = 0$ for any $j \geq d_0 + 1$, (A-2) leads to $x_j = \theta_j = 0$ and $\mathcal{R}(x, \varepsilon, j) = 1$, for any $j \geq d_0 + 1$. Moreover, Theorem 1 can be applied only for $d \leq d_0$ because f_{d_0+1} is not strictly positive and hence (A-3) fails. Consequently, $\varphi(x, \varepsilon) \sim f_{d_0}(x) V_{d_0}(\varepsilon)$ that is the usual first order approximation of the d_0 -dimensional process and f_{d_0} is the *intensity* of the SmBP of the process.

D.3 – Changing the basis. Let $\{\xi_j\}_{j=1}^\infty$ be an orthonormal basis of the Hilbert space arranged so that the sequence $\text{Var}(\langle X, \xi_j \rangle) = \lambda_j$ is sorted in descending order. Then Theorem 1 still holds.

3 Approximation results when d depends on ε

The goal of this section is to establish which conditions on X allow to simplify (9), to get

$$\varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon), \quad \text{as } \varepsilon \rightarrow 0.$$

In what follows, such result is achieved combining Theorem 1 and the limit behaviour of $\mathcal{R}(x, \varepsilon, d)$ which is strictly related to that of the real random variable $S(x, \varepsilon, d)$ defined by Equation (6). On the one hand, whether d and x are fixed, S diverges with ε tending to zero. On the other hand, if ε and x are fixed, the larger d the smaller S . Hence, one may wonder if it is possible to balance these two effects (as instance, tying the behaviour of d to that of ε) in order to have, for any x ,

$$\begin{cases} \mathcal{R}(x, \varepsilon, d) \rightarrow 1, \\ \varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon), \end{cases} \quad \varepsilon \rightarrow 0, \quad d(\varepsilon) \rightarrow \infty. \quad (10)$$

To do this let us consider the following limit behaviour of \mathcal{R} , as ε goes to zero and d diverges to infinity.

Proposition 2 Assume (A-2) and suppose that $\sum_{j \geq d+1} \lambda_j = o(1/d)$, as d goes to infinity. Then it is possible to choose $d = d(\varepsilon)$ so that it diverges to infinity as ε tends to zero and

$$d \sum_{j \geq d+1} \lambda_j = o(\varepsilon^2). \quad (11)$$

Moreover, as $\varepsilon \rightarrow 0$,

$$0 \leq 1 - \mathcal{R}(x, \varepsilon, d) \leq \frac{C_1(d+2)}{2\varepsilon^2} \sum_{j \geq d+1} \lambda_j = o(1). \quad (12)$$

Let us now consider the following inequality

$$|\varphi(x, \varepsilon) - f_d(x) V_d(\varepsilon)| \leq |\varphi(x, \varepsilon) - \varphi_d(x, \varepsilon)| + |\varphi_d(x, \varepsilon) - f_d(x) V_d(\varepsilon)|,$$

that, thanks to (8), (12) and to the fact that $0 < \mathcal{R} \leq 1$, leads to

$$\begin{aligned} \left| \frac{\varphi(x, \varepsilon)}{f_d(x) V_d(\varepsilon)} - 1 \right| &\leq C_2 \frac{\varepsilon^2}{2\lambda_d} \mathcal{R}(x, \varepsilon, d) + |\mathcal{R}(x, \varepsilon, d) - 1|, \\ &\leq C_2 \frac{\varepsilon^2}{2\lambda_d} + \frac{C_1(d+2)}{2\varepsilon^2} \sum_{j \geq d+1} \lambda_j. \end{aligned} \quad (13)$$

Thus, the wished result holds whenever the right-hand side vanishes as ε goes to zero, i.e. if there exists $d = d(\varepsilon)$ such that, at the same time the following two conditions hold

$$\varepsilon^2 = o(\lambda_d), \quad \text{and} \quad (d+2) \sum_{j \geq d+1} \lambda_j = o(\varepsilon^2). \quad (14)$$

In order to obtain (10), we combine conditions in (14) (plug the first in the second), and we get that eigenvalues must satisfy the *hyper-exponential* decay rate defined by

$$\frac{d \sum_{j \geq d+1} \lambda_j}{\lambda_d} = o(1), \quad \text{as } d \rightarrow \infty. \quad (15)$$

The latter highlights a trade-off between the approximation errors provided by Theorem 1 and Proposition 2. This trade-off is strictly related to a suitable balance between the d -th eigenvalue and the terms in the tail of the spectrum of the covariance operator.

It is worth noting that hyper-exponential decay of eigenvalues is a necessary condition to guarantee that the right-hand side of (13) vanishes. One may wonder, if it is sufficient as well, that is in other words, if it is possible to define $d = d(\varepsilon)$ so that the errors in (14) vanish at the same time as ε goes to zero. A positive answer to this is furnished by the following.

Theorem 3 Consider hypothesis of Theorem 1 and assume that eigenvalues decay hyper-exponentially. It is possible to choose $d = d(\varepsilon)$ so that, if $\varepsilon \rightarrow 0$, then $d \rightarrow \infty$ and

$$\varphi(x, \varepsilon) = f_d(x) V_d(\varepsilon) + o(f_d(x) V_d(\varepsilon)). \quad (16)$$

In what follows, in order to discuss assumptions and consequences of above result, some issues are developed.

D.4 – Again about the intensity of the SmBP. Because of the relation between d and ε , in general, (16) does not allow to define an *intensity* as commonly intended: anyway, being f_d the only term depending on x , it can be considered as a *surrogate intensity*.

The Gaussian processes or their suitable generalizations, provide examples for which f_d leads to define a genuine *intensity*. At first, consider a Gaussian process X , then for any $x \in \mathcal{L}_{[0,1]}^2$ and as ε goes to zero, $\varphi(x, \varepsilon) \sim \Psi_d(x) \cdot \mathcal{V}_d(\varepsilon)$; see D.1. When d tends to infinity, then $\Psi_d(x)$ tends to $\Psi(x) = \exp \left\{ -\sum_{j=1}^{\infty} x_j^2 / (2\lambda_j) \right\}$ for any $x \in \mathcal{L}_{[0,1]}^2$ and $\Psi(x)$ plays the role of the *intensity* of the small-ball probability at x . Note that, $\Psi(x)$ is not null if and only if x belongs to $RKHS(X)$, see (4). In particular, if we consider the Wiener process on $[0, 1]$, it can be proven, with arguments analogous to those in D.1 and for any smooth real function x on $[0, 1]$, that

$$\Psi(x) = \exp \left\{ -\frac{1}{2} \int_0^1 x'(t)^2 dt \right\}.$$

The latter is coherent with already known results; see, for instance, Li and Shao, 2001, Theorem 3.1.

Another situation in which an *intensity* for the SmBP can be also defined occurs whether the PCs are independent each one with density belonging to a subfamily of the exponential power (or generalized normal) distribution (see e.g. Box and Tiao, 1973), that is proportional to $\exp \left\{ -(|x_j|/\sqrt{\lambda_j})^q \right\}$, with $q \geq 2$. In this case, Ψ is given by

$$\Psi(x) = \exp \left\{ -\frac{1}{2} \sum_{j=1}^{\infty} \left(\frac{|x_j|}{\sqrt{\lambda_j}} \right)^q \right\}, \quad \text{for any } x \in \mathcal{L}_{[0,1]}^2$$

and, it is not null if x is in $H(q) = \left\{ x \in \mathcal{L}_{[0,1]}^2 : \sum_{j=1}^{\infty} (|x_j|/\sqrt{\lambda_j})^q < \infty \right\}$ that includes $RKHS(X)$ when $q \geq 2$.

D.5 – An example of hyper-exponential decay. Consider $\lambda_j = \exp\{-\beta j^\alpha\}$ with $\beta > 0$ and $\alpha > 1$. In this case, for any real number $n \geq 1$, it holds

$$\frac{d \sum_{j \geq d+1} \lambda_j}{\lambda_d} \leq \frac{d^n \sum_{j \geq d+1} \lambda_j}{\lambda_d} \rightarrow 0, \quad \text{as } d \rightarrow \infty. \quad (17)$$

In fact, some algebra and the Bernoulli inequality give

$$\begin{aligned} \frac{d^n \sum_{j \geq d+1} \lambda_j}{\lambda_d} &= d^n \left(\sum_{j \geq 1} \exp\{\beta d^\alpha (1 - (1 + j/d)^\alpha)\} \right) \\ &\leq d^n \left(\sum_{j \geq 1} \exp\{-\beta \alpha d^{\alpha-1} j\} \right). \end{aligned}$$

Since $\exp\{-\beta \alpha d^{\alpha-1} j\} \leq (j^2 d^{n+\delta})^{-1}$ eventually (with respect to d) holds for some positive δ and for each $j \in \mathbb{N}$, (17) is obtained.

3.1 Weakening the eigenvalues decay rate

If on the one hand, the factorization (16) provides the exact form of volumetric part (that is $V_d(\varepsilon)$, the volume of the d -dimensional ball of radius ε), on the other hand, it is obtained at the cost of the hyper-exponential eigenvalues decay (15). Weakening the eigenvalues decay rate, the exact form of the volumetric part no longer appears. In particular, we focus our interest on the following two further behaviours of the spectrum of Σ :

- “super-exponential”:

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j = o(1), \quad \text{as } d \rightarrow \infty.$$

or, equivalently,

$$\lambda_{d+1}/\lambda_d \rightarrow 0, \quad \text{as } d \rightarrow \infty. \quad (18)$$

- “exponential”: there exists a positive constant C so that

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \quad \text{for any } d \in \mathbb{N}. \quad (19)$$

It is possible to show that (15) \Rightarrow (18) \Rightarrow (19) but the contraries do not hold. For instance, for any $\alpha > 1$ and $\beta > 0$, $\lambda_j = \exp\{-\beta j\}$ decays exponentially but not super-exponentially, $\lambda_j = \exp\{-\beta j \ln(\ln(j))\}$ decays super-exponentially but not hyper-exponentially while $\lambda_j = \exp\{-\beta j^\alpha\}$ decays hyper-exponentially.

The following Theorem holds.

Theorem 4 *Under hypothesis of Theorem 1, as ε tends to zero, it is possible to choose $d = d(\varepsilon)$ diverging to infinity so that $\varphi(x, \varepsilon) \sim f_d(x) \phi(\varepsilon, d)$, where*

- *in the super-exponential case*

$$\phi(\varepsilon, d) = \exp \left\{ \frac{1}{2} d [\log(2\pi e \varepsilon^2) - \log(d) + o(1)] \right\},$$

- *in the exponential case*

$$\phi(\varepsilon, d) = \exp \left\{ \frac{1}{2} d [\log(2\pi e \varepsilon^2) - \log(d) + \delta(d, \alpha)] \right\},$$

where $\delta(\cdot, \cdot)$ is such that $\lim_{\alpha \rightarrow \infty} \limsup_{s \rightarrow \infty} \delta(s, \alpha) = 0$ and α is a parameter chosen so that $\lambda_d^{-1} \varepsilon^2 \leq \alpha^2$.

In other words, if the decay rate changes also the volume factor does. In particular, $f_d(x)$ preserves the role of a *surrogate intensity* whereas $\phi(\varepsilon, d)$ substitutes $V_d(\varepsilon)$ as volumetric term in the factorization. Observe that $\phi(\varepsilon, d)$ depends on terms (namely, $o(1)$ and $\delta(s, \alpha)$) that are implicitly defined and for which we just know the asymptotic behaviour. It is just the case to note that, in the exponential setting, Discussion D.4 about Gaussian and exponential power processes still holds with minor modifications.

D.6 – About slower eigenvalues decay rates. This theoretical problem is partially still open. In fact, a part from the Gaussian processes and, in particular, the Wiener

one (whose eigenvalues decay arithmetically but the intensity, evaluated at smooth x , can be defined as illustrated in D.1), to the best of our knowledge, there are no other attempts to provide asymptotic factorizations for the SmBP of processes whose eigenvalues decay slower than exponentially. Hence, at this stage, if no information about the probability law are available a solution is to go back to Theorem 1 in order to manage the dependence on x and ε in $\mathcal{R}(x, \varepsilon, d)$.

D.7 – Optimal basis. Although the factorization results in theorems 3 and 4 are stated by using the Karhunen–Loève (or PCA) basis, they hold for any orthonormal basis ordered according to the decreasing values of the variances of the projections, provided that they decay sufficiently fast. In particular, using the same notations as in D.3, if the sequence $\{\lambda_j\}_{j=1}^\infty$ has an exponential decay then Theorem 4 still holds and a *surrogate intensity* can be defined. Note that the variances obtained when one uses the PCA basis exhibit, by construction, the faster decay: in this sense the choice of this basis can be considered optimal.

4 Estimation of the surrogate intensity

Besides their theoretical interest, theorems 1, 3 and 4 turn to be useful from applications point of view as well. In fact, under suitable assumptions, they theoretically justify the use of f_d as a *surrogate intensity* for Hilbert-valued processes in statistical applications as done, for instance, within classification problems by Bongiorno and Goia (2016). This fact leads immediately to the main task of this section: to make the factorization results usable for practical purposes and, in particular, to introduce an estimator of the *surrogate intensity* f_d .

Consider a sample of random curves $\{X_i, i = 1, \dots, n\}$ which we suppose i.i.d. as X . In principle, if the sequence of eigenvalues $\{\xi_j\}_{j=1}^\infty$ was known, one should consider the empirical version of the vector of the first d principal components $\theta_i = (\theta_{1i}, \dots, \theta_{di})' \in \mathbb{R}^d$, with $\theta_{ji} = \langle X_i - \mathbb{E}[X_i], \xi_j \rangle$, and then introduce the classical kernel density estimate of f_d as follows:

$$f_{d,n}(\Pi_d x) = f_n(x) = \frac{1}{n} \sum_{i=1}^n K_{H_n}(\|\Pi_d(X_i - x)\|) \quad (20)$$

where $K_{H_n}(\mathbf{u}) = \det(H_n)^{-1/2} K(H_n^{-1/2}\mathbf{u})$, K is a kernel function and, $H_n = H_{nd}$ is a symmetric semi-definite positive $d \times d$ matrix (with an abuse of notations, we dropped the dependence on d). In practice, the equation (20) defines only a pseudo-estimate for f_d : indeed, the covariance operator Σ and then the sequence $\{\xi_j\}$ are unknown. Thus, to operationalize these pseudo-estimates it is necessary to consider the estimates $\hat{\theta}_i$ and $\hat{\Pi}_d$ of θ_i and Π_d respectively. In this view, consider the sample versions of μ_X and Σ , respectively:

$$\bar{X}_n(t) = \frac{1}{n} \sum_{i=1}^n X_i(t), \quad \text{and} \quad \hat{\Sigma}_n[\cdot] = \frac{1}{n} \sum_{i=1}^n \langle X_i - \bar{X}_n, \cdot \rangle (X_i - \bar{X}_n).$$

The eigenelements $\{\widehat{\lambda}_j, \widehat{\xi}_j\}_{j=1}^{\infty}$ of $\widehat{\Sigma}_n$ provide an estimation for $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ of Σ , as well as $\langle X_i - \overline{X}_n, \widehat{\xi}_j \rangle = \widehat{\theta}_{ji}$ estimates θ_{ji} (the asymptotic behaviour of these estimators has been widely studied; see e.g. Bosq, 2000). Thus, plugging these estimates (or the estimate of the eigen-projectors) in (20), we get the kernel density estimator:

$$\widehat{f}_{d,n}(\widehat{\Pi}_d x) = \widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n K_{H_n} \left(\left\| \widehat{\Pi}_d (X_i - x) \right\| \right), \quad \widehat{\Pi}_d x \in \mathbb{R}^d. \quad (21)$$

Since $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots \geq \widehat{\lambda}_n \geq 0 = \widehat{\lambda}_{n+1} = \dots$ one could choose $d = n$, but in practice this is not an appropriate choice: the curse of dimensionality problems in estimate a multivariate density combined with a bad estimation of the PCs associated to the smallest eigenelements would jeopardize the quality of estimation. Hence, a suitable dimension $d \ll n$ had to be identified. A first naive way consists in selecting the smallest d for which the Fraction of Explained Variance (defined as $\text{FEV}(d) = \sum_{j \leq d} \lambda_j / \sum_{j \geq 1} \lambda_j$) exceeds a fixed threshold. Anyway the problem of selecting the dimension d to be used in practice is still open and needs further developments that go beyond the scope of this paper.

If, from a computational point of view, the replacement of Π_d with $\widehat{\Pi}_d$ in (21) is a natural way to manage the problem of estimating the surrogate density in practice, one may wonder if that plug-in can influence the rate of convergence of the kernel estimator, or, in other words, if using \widehat{f}_n instead of f_n has no effect on this rate.

To answer this question, we study the behaviour of $\mathbb{E} \left[f_d(x) - \widehat{f}_n(x) \right]^2$ as n goes to infinity. For the sake of simplicity, we confine the study to the special case $H_n = h_n^2 I$ where I is the identity matrix, we assumed d fixed and independent of the observed data, and we suppose that

(B-1) the density $f_d(x)$ is positive and p times differentiable at $x \in \mathbb{R}^d$, with $p \geq 2$;

(B-2) the sequence of bandwidths h_n is such that:

$$h_n \rightarrow 0 \quad \text{and} \quad \frac{nh_n^d}{\log n} \rightarrow \infty \quad \text{as } n \rightarrow \infty;$$

(B-3) the kernel K is a Lipschitz, bounded, integrable density function with compact support $[0, 1]$;

(B-4) the process X satisfies the following condition: there exists two positive constants s and κ such that for all integer $m \geq 2$,

$$\mathbb{E} [\|X - x\|^m] \leq \frac{m!}{2} s \kappa^{m-2}.$$

The hypothesis (B-1), (B-2) and (B-3) are standard in the non-parametric framework, and $p \geq 2$ is required because of (A-3). Moreover, condition (B-4) holds for a wide family of processes including the Gaussian one.

Observe firstly that one can control the quadratic mean under study by intercalating the pseudo-estimator (21); in fact, thanks to the triangle inequality

$$\mathbb{E} \left[f_d(x) - \widehat{f}_n(x) \right]^2 \leq \mathbb{E} [f_d(x) - f_n(x)]^2 + \mathbb{E} \left[f_n(x) - \widehat{f}_n(x) \right]^2. \quad (22)$$

About the first term on the right-hand side of (22), it is well known in the literature (see for instance Wand and Jones, 1995) that under assumptions (B-1), ..., (B-4), and taking the optimal bandwidth

$$c_1 n^{-1/(2p+d)} \leq h_n \leq c_2 n^{-1/(2p+d)} \quad (23)$$

where c_1 and c_2 are two positive constants, one gets the minimax rate:

$$\mathbb{E} [f_d(x) - f_n(x)]^2 = O \left(n^{-2p/(2p+d)} \right)$$

uniformly in \mathbb{R}^d . Therefore, it is enough to control the second addend on the right-hand side of (22).

The following theorem states that, assuming a suitable degree of regularity for the density f_d depending on d , the rate of convergence in quadratic mean of $\widehat{f}_n(x)$ towards $f_n(x)$ is negligible with respect to the one of $f_n(x)$ towards $f_d(x)$. Thus, to use the estimated principal components instead of the empirical ones does not affect the rate of convergence.

Theorem 5 *Assume (B-1), ..., (B-4) with $p > 2 \vee 3d/2$ and consider the optimal bandwidth (23). Thus, as n goes to infinity,*

$$\mathbb{E} \left[f_n(x) - \widehat{f}_n(x) \right]^2 = o \left(n^{-2p/(2p+d)} \right),$$

uniformly in \mathbb{R}^d .

Formulation (21) requires that each random curve $X_i(t)$ is observed entirely in the continuum and without noise over $[0, 1]$. In practice, a discretization is inevitable as the curves are available only at discrete design points $\{\tau_{i,1}, \tau_{i,2}, \dots, \tau_{i,p_i}\}$, $\tau_{i,j} \in [0, 1]$, that are not necessarily the same for each i . Thus, it is necessary to introduce some numerical approximation to compute the estimates of the d principal components involved.

When each curve is observed without errors over the same fixed equispaced grid $\{\tau_1 = 0, \tau_2, \dots, \tau_{p-1}, \tau_p = 1\}$, with p sufficiently large, then one can replace simply integrals by summations: the empirical covariance operator is approximate by a matrix and its eigenelements are computed by standard numerical algorithms (see Rice and Silverman, 1991). This is the approach we follow in the simulations in Section 5 below.

A more general situation occurs when observed data are discretely sampled and corrupted by noise. In this case, one has the observed pairs $\{(\tau_{i,j}, Y_{i,j}), i = 1, \dots, n, j = 1, \dots, p_i\}$, where $Y_{i,j} = X_i(\tau_{i,j}) + \varepsilon_{ij}$ and the errors ε_{ij} are i.i.d. with zero mean and finite variance. If each $p_i \geq M_n$, where M_n is a suitable sequence tending to infinity with n (we refer to this case as *dense functional data*), a presmoothing process is run before to conduct PCA using the sample mean and covariance computed from the smoothed curves (see, for instance, Hall et al., 2006). In this case, under suitable hypothesis, the estimators of eigenelements are root- n consistent and first-order equivalent to the estimators obtained if curves were directly observed (see Hall et al., 2006, Theorem 3).

5 Finite sample performances in estimating the surrogate density

We illustrate now, through numerical examples, the feasibility of SmBP factorization approach by exploring how the proposed estimator works in a finite sample setting. We consider only two situations because of the difficulty in finding explicit expressions for the *intensity*. At first, we focus on a finite dimensional process for which the surrogate density is straightforward derived. After, we deal with the Wiener process that is one of the few infinite dimensional processes whose *intensity* can be derived, as already illustrated. In both cases, we study how the estimates behave varying the sample size and d . All simulations rest on the density estimator defined in (21), and are performed on a suitable grid of the d -dimensional factor space: the algorithms are implemented in R, and exploit the function `kde` in the package `ks` (see Duong, 2007).

5.1 Finite dimensional setting

Consider a one-dimensional random process whose trajectories are defined by

$$X(t) = a\sqrt{2/\pi} \sin(t), \quad t \in [0, \pi],$$

where a is a random variable with zero mean, unitary variance, density f_a and cumulative distribution function F_a . Given $x(t) = b\sqrt{2/\pi} \sin(t)$ with $b \in \mathbb{R}$, then, for any $\varepsilon > 0$, $\varphi(x, \varepsilon) = F_a(b + \varepsilon) - F_a(b - \varepsilon)$, and, as ε goes to zero, $\varphi(x, \varepsilon) \sim 2\varepsilon f_a(b)$. Such asymptotic is the same obtained from the SmBP factorization: since the first PC is $\theta = a$ and $x_1 = b$, it holds

$$\varphi(x, \varepsilon) \sim f_1(x_1) \varepsilon \pi^{1/2} / \Gamma(1/2 + 1) = 2f_a(b) \varepsilon, \quad \varepsilon \rightarrow 0,$$

with f_a being the *intensity* of the SmBP.

In this framework, f_a is compared with its estimates $\hat{f}_{1,n}$ from a sample of curves, for different $x(t)$, varying the nature of a and the sample size. In practice, set n and the distribution of a , we generated 1000 samples $\{X_i(t), i = 1, \dots, n\}$, i.i.d. as $X(t)$, (with $n = 50, 100, 200, 500, 1000$) where every curve is discretized over a mesh consisting on 100 equispaced points $\{t_j = (j - 1)\pi/99, j = 1, \dots, 100\}$. For each sample, we estimated the eigenfunction $\xi(t)$, the associated PC θ and its density via kernel procedure. Besides such samples, we built a set of curves $x^b(t) = b\sqrt{2/\pi} \sin(t)$ (discretized on the same grid as $X(t)$), where b is a suitable increasing sequence of real values. The estimated density $\hat{f}_{1,n}$ is then evaluated at the points $\hat{x}_1^b = \langle x^b(t), \hat{\xi}(t) \rangle$ and compared with the true values $f_a(b)$ in term of relative mean square prediction error (RMSEP = $\sum_b [\hat{f}_{1,n}(\hat{x}_1^b) - f_a(b)]^2 / \sum_b f_a^2(b)$) over the 1000 replications. Moreover it is also possible investigate for what values b the estimate of the surrogate density is better, by using the absolute percentage error (APE = $|\hat{f}_{1,n}(\hat{x}_1^b) - f_a(b)| / f_a(b)$).

In the experiment we take a distributed as: 1) standard Gaussian (that is, $a \sim \mathcal{N}(0, 1)$), 2) a standardized Student t with 5 df (that is, $a \sim t(5) / \sqrt{5/3}$), 3) a standardized Chi-square distribution with 8 df (that is, $a \sim (\chi^2(8) - 8) / 4$). About b , we

n	$N(0, 1)$		$t(5) / \sqrt{5/3}$		$(\chi^2(8) - 8) / 4$	
	Mean	Std.	Mean	Std.	Mean	Std.
50	3.235	(2.681)	5.921	(2.557)	4.081	(2.842)
100	1.860	(1.444)	4.775	(1.503)	2.401	(1.619)
200	1.091	(0.824)	4.138	(0.878)	1.422	(0.887)
500	0.546	(0.355)	3.737	(0.477)	0.753	(0.443)
1000	0.330	(0.220)	3.606	(0.327)	0.453	(0.233)

Table 1: Mean and standard deviation of RMSEP (Results $\times 10^{-2}$) for Gaussian, t and χ^2 distributions, computed over 1000 Monte Carlo replications varying the sample size n .

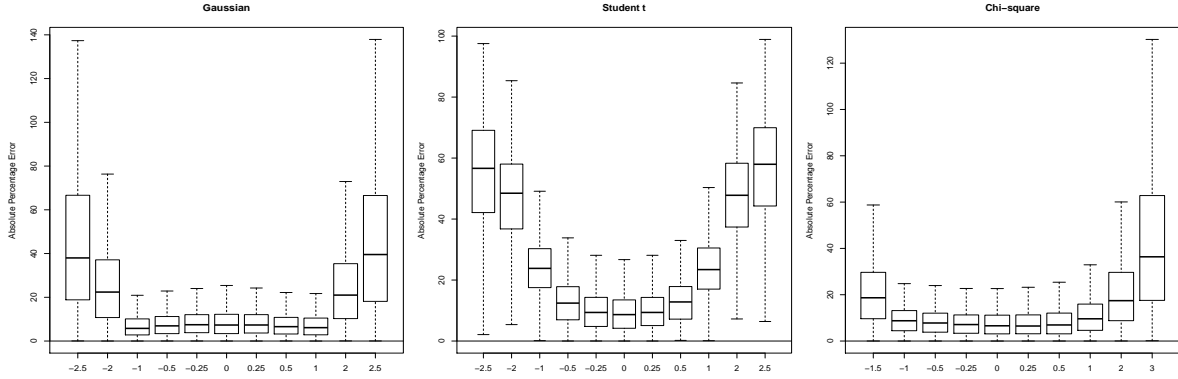


Figure 1: Absolute percentage errors in estimating $f_a(b)$ varying b for Normal, t and χ^2 distributions respectively.

used sequences consisting of 160 equispaced points, over the interval $[-4, 4]$ for the distributions 1) and 2), and $[-2, 6]$ for the asymmetric distribution 3).

The MSE ($\times 10^{-2}$) obtained under the different experimental conditions are collected in Table 1. As expected, results improve as the sample size increases: that is due both to the better estimates of projections $\hat{\theta}$ and \hat{x}^b and to the better performances of the kernel estimator. On the other hand, differences due to the shape of distributions occur: long tails and asymmetries produce a deterioration in estimates.

The APE for some selected values b when $n = 200$ are reproduced in Figure 1. As one can expect, the quality of estimate worsens at the edges of the distributions, when b is rather far from zero. This fact is connected to the limitations of kernel density estimator in evaluate the tails of distributions.

5.2 Infinite dimensional setting

In this second experiment, we deal with an infinite dimensional setting in order to study how the estimation of the *intensity* of the SmBP behaves according to the sample size and the dimensional parameter d . To do this, keeping in mind D.1, let us consider a Wiener process X on $[0, 1]$ and the smooth function $x(t) = \sum_{j=1}^{d_0} b_j \xi_j(t)$, with, for the

n	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$
50	3.361 (2.506)	7.197 (3.728)	13.529 (7.338)	22.029 (12.052)	31.900 (15.871)	42.051 (19.159)
100	1.948 (1.198)	4.821 (2.585)	9.471 (5.535)	15.857 (8.711)	23.988 (12.272)	33.729 (15.871)
200	1.158 (0.719)	3.143 (1.601)	6.638 (3.773)	11.514 (6.295)	17.894 (9.476)	25.514 (13.098)
500	0.574 (0.334)	1.775 (0.932)	4.168 (2.363)	7.766 (4.232)	12.956 (6.884)	18.988 (9.284)
1000	0.345 (0.187)	1.149 (0.625)	2.822 (1.636)	5.858 (3.126)	10.091 (5.425)	15.292 (7.651)

Table 2: Mean and standard deviation (in parentheses) of RMSEP (Results $\times 10^{-2}$) for Wiener process, computed over 1000 Monte Carlo replications varying the sample size n and the dimension d .

sake of simplicity, $d_0 = 1$, that is

$$x(t) = b \frac{2\sqrt{2}}{\pi} \sin\left(\frac{\pi t}{2}\right), \quad t \in [0, 1], \quad (24)$$

where $b \in \mathbb{R}$, so that the *intensity* is $\Psi_{d_0}(x) = \exp\{-b^2/2\}$.

The experiment follows a similar route as in 5.1. In a first step, we generated 1000 samples $\{X_i(t), i = 1, \dots, n\}$, (with $n = 50, 100, 200, 500, 1000$) where every curve is discretized over 100 equispaced points $\mathcal{G} = \{t_j = (j-1)/99, j = 1, \dots, 100\}$, and 160 fixed curves $x^b(t)$ generated according to (24) and discretized over \mathcal{G} (b is an increasing sequence of equispaced points, over the interval $[-4, 4]$). In a second step, for each sample, once empirical eigenfunctions $\hat{\xi}_j(t)$ are obtained, we estimated f_d (with $d = 1, \dots, 6$) and we computed them at $(\hat{x}_1^b, \dots, \hat{x}_d^b)'$ where $\hat{x}_j^b = \langle x^b(t), \hat{\xi}_j(t) \rangle$. Finally, we compared the estimated surrogate density with the true one in term of relative mean square prediction error (MSEP) over the 1000 replications. The obtained results, varying n and d are reported in Table 2.

As a general comment, one can observe that for each d the MSPE reduces (both in mean and in variability) increasing n , whereas, for each n , the MSPE increases (both in mean and in variability) with d . This shows how the curse of dimensionality interferes in the kernel estimation procedure as soon as the dimension d exceeds one. To perceive the relation between d and n , one has to read the table in diagonal sense: it is possible to use large d at the cost of use large samples. For instance, similar very good results (at around 3%) are possible using $n = 50$ and $d = 1$, or $n = 200$ and $d = 2$, or when $n = 1000$ and $d = 3$. On the other hand, results benefit from the fact that the spectrum of the process is rather concentrate. In fact, the Fraction of Explained Variance (defined as $\text{FEV}(d) = \sum_{j \leq d} \lambda_j / \sum_{j \geq 1} \lambda_j$) are: $\text{FEV}(1) = 0.811$, $\text{FEV}(2) = 0.901$, $\text{FEV}(3) = 0.933$, $\text{FEV}(4) = 0.950$, $\text{FEV}(5) = 0.960$ and $\text{FEV}(6) = 0.966$. Hence, good estimates for the surrogate density are already possible with $d = 1$ or $d = 2$, also for medium size samples. In that sense, this experiment gives an empirical evidence on the use of FEV in select the dimensional parameter d , conscious of the fact that large d need large n in order to have better estimates.

6 Proofs

This section collects proofs of results exposed above.

6.1 Proof of Theorem 1

We are interested in the asymptotic behaviour, whenever ε tends to zero, of the SmBP of the process X , that is

$$\begin{aligned}\varphi(x, \varepsilon) &= \mathbb{P}(\|X - x\| \leq \varepsilon) = \mathbb{P}(\|X - x\|^2 \leq \varepsilon^2) \\ &= \mathbb{P}\left(\sum_{j=1}^{+\infty} \langle X - x, \xi_j \rangle^2 \leq \varepsilon^2\right) = \mathbb{P}\left(\sum_{j=1}^{+\infty} (\theta_j - x_j)^2 \leq \varepsilon^2\right), \quad \text{as } \varepsilon \rightarrow 0\end{aligned}$$

Let $S_1 = \sum_{j \leq d} (\theta_j - x_j)^2$ and $S = \frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2$ be the truncated series and the scaled version of the remainder respectively. Thus, the SmBP is

$$\begin{aligned}\varphi(x, \varepsilon) &= \mathbb{P}(S_1 + \varepsilon^2 S \leq \varepsilon^2) = \mathbb{P}(S_1 \leq \varepsilon^2 (1 - S)) \\ &= \mathbb{P}(\{S_1 \leq \varepsilon^2 (1 - S)\} \cap \{S \geq 1\}) + \mathbb{P}(\{S_1 \leq \varepsilon^2 (1 - S)\} \cap \{0 \leq S < 1\}) \\ &= \mathbb{P}(\{S_1 \leq \varepsilon^2 (1 - S)\} \cap \{0 \leq S < 1\}) \\ &= \int_0^1 \varphi(s|x, \varepsilon, d) dG(s)\end{aligned}\tag{25}$$

where G is the cumulative distribution function of S . At first, for any $s \in (0, 1)$, let us consider $\varphi(s|x, \varepsilon, d)$, that is the SmBP about $\Pi_d x$ of the process $\Pi_d X$ in the space spanned by $\{\xi_j\}_{j \leq d}$. In terms of $f_d(\cdot)$, the probability density function of $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_d)'$, it can be written as

$$\varphi(s|x, \varepsilon, d) = \int_{D^x} f_d(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta},$$

where $D = D^x = \{\boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j \leq d} (\vartheta_j - x_j)^2 \leq \varepsilon^2 (1 - s)\}$ is a d -dimensional ball centered about $\Pi_d x = (x_1, \dots, x_d)$ with radius $\varepsilon \sqrt{1 - s}$. Now, consider the Taylor expansion of $f = f_d$ about $\Pi x = \Pi_d x$,

$$\begin{aligned}f(\boldsymbol{\vartheta}) &= f(x_1, \dots, x_d) + \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \dots, x_d) \rangle \\ &\quad + \frac{1}{2} (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t) (\boldsymbol{\vartheta} - \Pi x),\end{aligned}$$

for some $t \in (0, 1)$ and with H_f denoting the Hessian matrix of f . (In general, t depends on $\boldsymbol{\vartheta} - \Pi x$, but we are not interested in the actual value of it because the boundedness of the second derivatives of f allows us to drop, in what follows, those

terms depending on t). Then we can write

$$\begin{aligned}
\varphi(s|x, \varepsilon, d) &= \int_D \left(f(x_1, \dots, x_d) + \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \dots, x_d) \rangle \right. \\
&\quad \left. + \frac{1}{2} (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t) (\boldsymbol{\vartheta} - \Pi x) \right) d\boldsymbol{\vartheta} \\
&= f(x_1, \dots, x_d) \int_D d\boldsymbol{\vartheta} + \int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \dots, x_d) \rangle d\boldsymbol{\vartheta} \\
&\quad + \frac{1}{2} \int_D (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t) (\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \\
&= f(x_1, \dots, x_d) I + \frac{1}{2} \int_D (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t) (\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \quad (26)
\end{aligned}$$

where $I = I(s, \varepsilon, d)$ denotes the volume of D that is

$$I = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2 + 1)} (1 - s)^{d/2}$$

and, the addend $\int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \dots, x_d) \rangle d\boldsymbol{\vartheta}$ is null since $\langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \dots, x_d) \rangle$ is a linear functional integrated over the symmetric – with respect to the center (x_1, \dots, x_d) – domain D . Thus from (26), thanks to: the boundedness of second derivatives (5), the fact that symmetry arguments lead to $\int_D (\vartheta_i - x_i)(\vartheta_j - x_j) d\boldsymbol{\vartheta} = 0$ for $i \neq j$ and monotonicity of eigenvalues, it follows

$$\begin{aligned}
|\varphi(s|x, \varepsilon, d) - f(x_1, \dots, x_d) I| &= \\
&= \left| \frac{1}{2} \int_D \sum_{i \leq d} \sum_{j \leq d} (\vartheta_i - x_i)(\vartheta_j - x_j) \frac{\partial^2 f}{\partial \vartheta_i \partial \vartheta_j} (\Pi x + (\boldsymbol{\vartheta} - \Pi x)t) d\boldsymbol{\vartheta} \right| \\
&\leq \frac{1}{2} C_2 f(x_1, \dots, x_d) \left| \sum_{i \leq d} \sum_{j \leq d} \int_D \frac{(\vartheta_i - x_i)(\vartheta_j - x_j)}{\sqrt{\lambda_i} \sqrt{\lambda_j}} d\boldsymbol{\vartheta} \right| \\
&= \frac{1}{2} C_2 f(x_1, \dots, x_d) \int_D \sum_{j \leq d} \frac{(\vartheta_j - x_j)^2}{\lambda_j} d\boldsymbol{\vartheta} \\
&\leq \frac{C_2}{2\lambda_d} f(x_1, \dots, x_d) \int_D \sum_{j \leq d} (\vartheta_j - x_j)^2 d\boldsymbol{\vartheta}.
\end{aligned}$$

Note that

$$\int_D \sum_{j \leq d} (\vartheta_j - x_j)^2 d\boldsymbol{\vartheta} = \int_{\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d}^2 \leq \varepsilon^2(1-s)} \|\boldsymbol{\vartheta}\|_{\mathbb{R}^d}^2 d\boldsymbol{\vartheta}$$

whose integrand is a radial function (i.e. a map $H : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $H(\boldsymbol{\vartheta}) = h(\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d})$ with $h : \mathbb{R} \rightarrow \mathbb{R}$), for which the following identity applies

$$\int_{\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d} \leq R} H(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \omega_{d-1} \int_0^R h(\rho) \rho^{d-1} d\rho,$$

where ω_{d-1} denotes the surface area of the sphere of radius 1 in \mathbb{R}^d . Hence

$$\int_{\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d}^2 \leq \varepsilon^2(1-s)} \|\boldsymbol{\vartheta}\|_{\mathbb{R}^d}^2 d\boldsymbol{\vartheta} = \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_0^{\varepsilon\sqrt{1-s}} \rho^{d+1} d\rho = \frac{d}{(d+2)} I\varepsilon^2(1-s) \leq I\varepsilon^2,$$

where the latter inequality follows from the fact that $s \in [0, 1)$. This leads to

$$|\varphi(s|x, \varepsilon, d) - f(x_1, \dots, x_d)I| \leq C_2 \frac{\varepsilon^2 I}{2\lambda_d} f(x_1, \dots, x_d). \quad (27)$$

Come back to the SmBP (25),

$$\varphi(x, \varepsilon) = \int_0^1 f(x_1, \dots, x_d) IdG(s) + \int_0^1 (\varphi(s|x, \varepsilon, d) - f(x_1, \dots, x_d)I) dG(s), \quad (28)$$

and note that, thanks to (27) and because d is fixed, the second addend in the right-hand side of (28) is infinitesimal with respect to the first addend

$$\begin{aligned} & \left| \frac{\int_0^1 (\varphi(s|x, \varepsilon, d) - f(x_1, \dots, x_d)I) dG(s)}{\int_0^1 f(x_1, \dots, x_d) IdG(s)} \right| \leq \\ & \leq \left| \frac{C_2 \frac{\varepsilon^2}{2\lambda_d} f(x_1, \dots, x_d) \int_0^1 IdG(s)}{f(x_1, \dots, x_d) \int_0^1 IdG(s)} \right| = C_2 \frac{\varepsilon^2}{2\lambda_d}. \end{aligned}$$

Noting that

$$\int_0^1 I(s, \varepsilon, d) dG(s) = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2 + 1)} \mathbb{E} \left[(1 - S)^{d/2} \mathbb{I}_{\{S \leq 1\}} \right],$$

we obtain

$$|\varphi(x, \varepsilon) - \varphi_d(x, \varepsilon)| \leq C_2 \frac{\varepsilon^2}{2\lambda_d} \varphi_d(x, \varepsilon) \quad (8)$$

where,

$$\varphi_d(x, \varepsilon) = f(x_1, \dots, x_d) \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2 + 1)} \mathbb{E} \left[(1 - S)^{d/2} \mathbb{I}_{\{S \leq 1\}} \right]. \quad (7)$$

Thus, since d is fixed, as ε tends to zero,

$$\varphi(x, \varepsilon) = \int_0^1 \varphi(s|x, \varepsilon, d) dG(s) = \varphi_d(x, \varepsilon) + o\left(\frac{\varphi_d(x, \varepsilon)}{f(x_1, \dots, x_d)}\right)$$

or, equivalently, $\varphi(x, \varepsilon) \sim \varphi_d(x, \varepsilon)$ that concludes the proof.

6.2 Proofs of Proposition 2, and theorems 3 and 4

To prove Proposition 2 we need the following Lemma.

Lemma 6 Assume (A-1) and (A-2). Then, it is possible to choose $d = d(\varepsilon)$ so that it diverges to infinity as ε tends to zero and

$$\sum_{j \geq d+1} \lambda_j = o(\varepsilon^2). \quad (29)$$

Moreover, as $\varepsilon \rightarrow 0$, $S(x, \varepsilon, d) \rightarrow 0$, where the convergence holds almost surely, in the L^1 norm and hence in probability.

Proof. A possible choice for $d = d(\varepsilon)$ satisfying (29) can be, for a fixed $\delta > 0$, as follow

$$d = \min \left\{ k \in \mathbb{N} : \sum_{j \geq k+1} \lambda_j \leq \varepsilon^{2+\delta} \right\}, \quad \text{for any } \varepsilon > 0.$$

Such a minimum is well defined since eigenvalues series is convergent.

Let us prove that S converges to zero in probability. For any $k > 0$, by Markov inequality and, thanks to Equation (3),

$$\begin{aligned} \mathbb{P}(|S| > k) &= \mathbb{P}(S > k) = \mathbb{P}\left(\frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2 > k\right) \\ &\leq \frac{\mathbb{E}\left[\frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2\right]}{k^2} \leq \frac{C_1 \sum_{j \geq d+1} \lambda_j}{k^2 \varepsilon^2}. \end{aligned} \quad (30)$$

Thanks to (29) we get the convergence in probability. Since $S = S(x, \varepsilon, d)$ is non-increasing when d increases,

$$\mathbb{P}\left(\sup_{j \geq d+1} |S(x, \varepsilon, j) - 0| \geq k\right) = \mathbb{P}(S(x, \varepsilon, d+1) \geq k)$$

holds for any $k > 0$ and any x . This fact, together with (30), guarantees the almost sure convergence of S to zero (e.g. Shiriyayev, 1984, Theorem 10.3.1) as ε tends to zero. Moreover, the monotone convergence theorem guarantees the L^1 convergence. ■

Proof of Proposition 2. Note that if $d(\varepsilon)$ satisfies (11), then (29) and Lemma 6 hold. For a fixed $\delta > 0$, a possible choice of such $d = d(\varepsilon)$ can be

$$d = \min \left\{ k \in \mathbb{N} : k \sum_{j \geq k+1} \lambda_j \leq \varepsilon^{2+\delta} \right\},$$

where the minimum is achieved thanks to the eigenvalues hyperbolic decay assumption. At this stage, note that

$$0 < \mathbb{E}\left[(1 - S)^{d/2} \mathbb{I}_{\{S < 1\}}\right] \leq 1$$

then, after some algebra, thanks to Bernoulli inequality (i.e. $(1+s)^r \geq 1+rs$ for $s \geq -1$ and $r \in \mathbb{R} \setminus (0, 1)$), Markov inequality and Assumption (3), we have (for any $d \geq 2$)

$$\begin{aligned} 0 &\leq 1 - \mathbb{E} \left[(1 - S)^{d/2} \mathbb{I}_{\{S < 1\}} \right] \leq 1 - \mathbb{E} \left[\left(1 - \frac{d}{2} S \right) \mathbb{I}_{\{S < 1\}} \right] \\ &\leq \mathbb{P}(S \geq 1) + \mathbb{E} \left[\frac{d}{2} S \mathbb{I}_{\{S < 1\}} \right] \leq \mathbb{E} \left[\frac{(d+2)}{2\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2 \right] \leq \frac{C_1(d+2)}{2\varepsilon^2} \sum_{j \geq d+1} \lambda_j. \end{aligned}$$

Choosing d according to (11) the thesis follows. ■

Proof of Theorem 3. Thanks to hyper-exponentiality (15), there exists $d_0 \in \mathbb{N}$ so that for any $d \geq d_0$

$$d \sum_{j \geq d+1} \lambda_j < \lambda_d.$$

Moreover, there exist $\delta_1, \delta_2 \in (0, 1)$ (depending on d) for which, for any $d \geq d_0$

$$0 \leq d \sum_{j \geq d+1} \lambda_j \leq b(d, \{\lambda_j\}_{j \geq d+1}, \delta_1) < B(d, \{\lambda_j\}_{j \leq d}, \delta_2) \leq \lambda_d, \quad (31)$$

where

$$b(d, \{\lambda_j\}_{j \geq d+1}, \delta_1) = \left(d \sum_{j \geq d+1} \lambda_j \right)^{1-\delta_1}, \quad B(d, \{\lambda_j\}_{j \leq d}, \delta_2) = \lambda_d^{1-\delta_2}.$$

As instance, for a given $d \geq d_0$, fix $\delta_1 \in (0, 1)$ and solve (31) with respect to δ_2 , that is $\delta_2 \in (\min \{0, \beta(\delta_1)\}, 1)$ where $\beta(\delta_1) = 1 - (1 - \delta_1) \ln \left(d \sum_{j \geq d+1} \lambda_j \right) / \ln(\lambda_d)$. As a consequence, for any $\varepsilon > 0$ and for such a choice of δ_1, δ_2 , the following minimum is well-defined

$$d(\varepsilon) = \min \{ k \in \mathbb{N} : b(k, \{\lambda_j\}_{j \geq k+1}, \delta_1) \leq \varepsilon^2 \leq B(k, \{\lambda_j\}_{j \leq k}, \delta_2) \}.$$

This guarantees that the right-hand side of (13) vanishes as ε goes to zero. ■

To prove Theorem 4 we need the following Lemma.

Lemma 7 Assume (A-1) and (A-2). Then, as $\varepsilon \rightarrow 0$,

$$\mathcal{R}(x, \varepsilon, d)^{2/d} \rightarrow 1, \quad \text{or,} \quad \log(\mathcal{R}(x, \varepsilon, d)) = o(d). \quad (32)$$

Proof. Jensen inequality for concave functions (i.e. $\mathbb{E}[f(g)] \leq f(\mathbb{E}[g])$ if f is a concave function) guarantees that

$$\begin{aligned} \mathbb{E} \left[((1 - S) \mathbb{I}_{\{S < 1\}})^{\frac{d}{2}} \right] &= \mathbb{E} \left[((1 - S) \mathbb{I}_{\{S < 1\}})^{\frac{d+1}{2} \frac{d}{d+1}} \right] \\ &\leq \left\{ \mathbb{E} \left[((1 - S) \mathbb{I}_{\{S < 1\}})^{\frac{d+1}{2}} \right] \right\}^{\frac{d}{d+1}}, \end{aligned}$$

noting that $S(x, \varepsilon, d+1) =: S_{d+1} \leq S_d := S(x, \varepsilon, d)$ and $\mathbb{I}_{\{S_d < 1\}} \leq \mathbb{I}_{\{S_{d+1} < 1\}}$, then

$$\mathbb{E} \left[((1 - S_d) \mathbb{I}_{\{S_d < 1\}})^{\frac{d}{2}} \right] \leq \left\{ \mathbb{E} \left[((1 - S_{d+1}) \mathbb{I}_{\{S_{d+1} < 1\}})^{\frac{d+1}{2}} \right] \right\}^{\frac{d}{d+1}}.$$

The latter guarantees that $\mathbb{E} \left[(1 - S)^{d/2} \mathbb{I}_{\{S < 1\}} \right]^{2/d}$ is a non-decreasing monotone sequence with respect to d whose values are in $(0, 1]$ and eventually bounded away from zero. ■

Proof of Theorem 4. Given results in Theorem 1, thesis holds using same arguments as in (Delaigle and Hall, 2010, Proof of Theorem 4.2.): the idea is to combine together (32), the Stirling expansion of the Gamma function in V_d and the (super-)exponential eigenvalues decay. ■

6.3 Proof of Theorem 5

In what follows, as did in Section 4, we simplify the notations dropping the dependence on d for the density estimators f_n and \hat{f}_n . Moreover, C denotes a general positive constant. The proof of Theorem 5 uses similar arguments as in Biau and Mas (2012). Since $H_n = h_n^2 I$, it holds $K_{H_n}(u) = h_n^{-d} K(u)$. Consider

$$S_n(x) = \sum_{i=1}^n K \left(\frac{\|\Pi_d(X_i - x)\|}{h_n} \right), \quad \hat{S}_n(x) = \sum_{i=1}^n K \left(\frac{\|\hat{\Pi}_d(X_i - x)\|}{h_n} \right),$$

then the pseudo-estimator and the estimator are given by

$$f_n(x) = \frac{S_n(x)}{nh_n^d}, \quad \hat{f}_n(x) = \frac{\hat{S}_n(x)}{nh_n^d},$$

and, hence,

$$\mathbb{E} \left[f_n(x) - \hat{f}_n(x) \right]^2 = \frac{1}{(nh_n^d)^2} \mathbb{E} \left[S_n(x) - \hat{S}_n(x) \right]^2.$$

Set $V_i = \|\Pi_d(X_i - x)\|$, $\hat{V}_i = \|\hat{\Pi}_d(X_i - x)\|$, consider the events

$$A_i = \{V_i \leq h_n\}, \quad B_i = \{\hat{V}_i \leq h_n\},$$

then we have the decomposition

$$\begin{aligned} S_n(x) - \hat{S}_n(x) &= \sum_{i=1}^n \left[K \left(\frac{V_i}{h_n} \right) - K \left(\frac{\hat{V}_i}{h_n} \right) \right] \mathbb{I}_{A_i \cap B_i} + \\ &\quad + \sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \bar{B}_i} - \sum_{i=1}^n K \left(\frac{\hat{V}_i}{h_n} \right) \mathbb{I}_{\bar{A}_i \cap B_i}. \end{aligned}$$

Since $(a + b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned} \mathbb{E} \left[S_n(x) - \widehat{S}_n(x) \right]^2 &\leq 2\mathbb{E} \left[\sum_{i=1}^n \left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right) \right) \mathbb{I}_{A_i \cap B_i} \right]^2 \\ &\quad + 2\mathbb{E} \left[\left(\sum_{i=1}^n K\left(\frac{V_i}{h_n}\right) \mathbb{I}_{A_i \cap B_i} \right)^2 + \left(\sum_{i=1}^n K\left(\frac{\widehat{V}_i}{h_n}\right) \mathbb{I}_{\bar{A}_i \cap B_i} \right)^2 \right]. \end{aligned} \quad (33)$$

Consider now the first addend in the right-hand side of (33): Assumption (B-3) and the fact that $|V_i - \widehat{V}_i| \leq \|\Pi_d - \widehat{\Pi}_d\|_\infty \|X_i - x\|$, where $\|\cdot\|_\infty$ denotes the operator norm, lead to

$$\mathbb{E} \left[\sum_{i=1}^n \left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right) \right) \mathbb{I}_{A_i \cap B_i} \right]^2 \leq C\mathbb{E} \left[\|\Pi_d - \widehat{\Pi}_d\|_\infty^2 \sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} \right]^2.$$

Thanks to the Cauchy-Schwartz inequality we control the previous bound by

$$C\mathbb{E} \left[\|\Pi_d - \widehat{\Pi}_d\|_\infty^2 \right] \mathbb{E} \left[\left(\sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} \right)^2 \right]. \quad (34)$$

About the first factor in (34), Biau and Mas, 2012, Theorem 2.1 (ii) established that

$$\mathbb{E} \left[\|\Pi_d - \widehat{\Pi}_d\|_\infty^2 \right] = O\left(\frac{1}{n}\right). \quad (35)$$

Consider now the second term in (34). Thanks to the Chebyshev's algebraic inequality (see, for instance, Mitrović et al., 1993, page 243) and since $\mathbb{E}[\mathbb{I}_{A_i \cap B_i}] \leq \mathbb{E}[\mathbb{I}_{A_i}]$, for any $k \geq 1$ it holds

$$\mathbb{E} \left[\|X - x\|^k \mathbb{I}_{A_i \cap B_i} \right] \leq \mathbb{E} \left[\|X - x\|^k \right] \mathbb{E}[\mathbb{I}_{A_i}].$$

The fact that $\mathbb{E}[\mathbb{I}_{A_i}] \sim h_n^d$ and Assumption (B-4) give

$$\mathbb{E} \left[\|X - x\|^k \mathbb{I}_{A_i \cap B_i} \right] \leq C \frac{k!}{2} b^{k-2} h_n^d,$$

with $b > 0$. Hence, the Bernstein inequality (see e.g. Massart, 2007) can be applied: for any $M > 0$,

$$\mathbb{P} \left(\left| \sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} - \mathbb{E} \left[\sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} \right] \right| \geq Mnh^d \right) \leq \exp(-CM^2nh^d).$$

This result together with the Borel-Cantelli lemma lead to:

$$\sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} \leq Cnh^d \quad a.s.$$

and therefore,

$$\mathbb{E} \left[\left(\sum_{i=1}^n \|X_i - x\| \mathbb{I}_{A_i \cap B_i} \right)^2 \right] \leq C n^2 h^{2d}. \quad (36)$$

Finally, combining results (35) and (36), we obtain:

$$\frac{1}{(n h_n^d)^2} \mathbb{E} \left[\sum_{i=1}^n \left(K \left(\frac{V_i}{h_n} \right) - K \left(\frac{\widehat{V}_i}{h_n} \right) \right) \mathbb{I}_{A_i \cap B_i} \right]^2 \leq C \frac{1}{n h_n^2}. \quad (37)$$

Consider now the second addend in the right-hand side of (33). We only look at

$$\mathbb{E} \left[\sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \overline{B}_i} \right]^2, \quad (38)$$

because the behaviour of the other addend is similar. Define the sequence κ_n so that $\kappa_n \rightarrow 0$ as $n \rightarrow \infty$, the following inclusions hold:

$$\begin{aligned} A_i \cap \overline{B}_i &= \{V_i \leq h_n\} \cap \{\widehat{V}_i > h_n\} \\ &= (\{h_n(1 - \kappa_n) < V_i \leq h_n\} \cup \{V_i \leq h_n(1 - \kappa_n)\}) \cap \{\widehat{V}_i - V_i > h_n - V_i\} \\ &\subseteq \{h_n(1 - \kappa_n) < V_i \leq h_n\} \cup \{V_i \leq h_n(1 - \kappa_n), \widehat{V}_i - V_i > h_n - V_i\} \\ &\subseteq \{h_n(1 - \kappa_n) < V_i \leq h_n\} \cup \{\widehat{V}_i - V_i > \kappa_n h_n\}. \end{aligned}$$

The latter inclusion and Assumption (B-3) allow to control (38) by

$$\mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{A_i \cap \overline{B}_i} \right]^2 \leq 2 \mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{\{h_n(1 - \kappa_n) < V_i \leq h_n\}} \right]^2 + 2 \mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{\{\|\widehat{\Pi}_d - \Pi_d\| \|X_i - x\| > C \kappa_n h_n\}} \right]^2. \quad (39)$$

About the first term in the right-hand side of the latter, the Cauchy-Schwartz inequality gives

$$\mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{\{h_n(1 - \kappa_n) < V_i \leq h_n\}} \right]^2 \leq n^2 \mathbb{P}(h_n(1 - \kappa_n) < V \leq h_n).$$

Since $\mathbb{P}(h_n(1 - \kappa_n) < V \leq h_n) \sim h_n^d (1 - (1 - \kappa_n)^d)$, performing a first order Taylor expansion of $(1 - \kappa_n)^d$ in $\kappa_n = 0$, we get asymptotically

$$\mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{\{h_n(1 - \kappa_n) < V_i \leq h_n\}} \right]^2 \leq C n^2 h_n^d \kappa_n.$$

Similarly, for what concerns the other addend in the right-hand side of (39), we have

$$\mathbb{E} \left[\sum_{i=1}^n \mathbb{I}_{\{\|\widehat{\Pi}_d - \Pi_d\| \|X_i - x\| > C \kappa_n h_n\}} \right]^2 \leq n^2 \mathbb{P}(\|\widehat{\Pi}_d - \Pi_d\| \|X - x\| > C \kappa_n h_n).$$

Thanks to the Markov inequality, Biau and Mas, 2012, Theorem 2.1 (iii) and Assumption (B-4), it follows

$$\mathbb{P} \left(\left\| \widehat{\Pi}_d - \Pi_d \right\| \|X - x\| > C\kappa_n h_n \right) = O \left(\frac{1}{n^{1/2} h_n \kappa_n} \right).$$

Combining the previous results we obtain:

$$\frac{1}{(nh_n^d)^2} \mathbb{E} \left[\left(\sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \overline{B}_i} \right)^2 \right] = O \left(\frac{\kappa_n}{h_n^d} \right) + O \left(\frac{1}{n^{1/2} h_n \kappa_n} \right).$$

If we choose $\kappa_n = (n^{5/2} h_n^{2d})^{-1/2}$ with $n^{5/4} h_n^d \rightarrow \infty$, as $n \rightarrow \infty$, we obtain:

$$\mathbb{E} \left[\left(\sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \overline{B}_i} \right)^2 + \left(\sum_{i=1}^n K \left(\frac{\widehat{V}_i}{h_n} \right) \mathbb{I}_{\overline{A}_i \cap B_i} \right)^2 \right] \leq C \frac{1}{n^{5/4} h_n^{2d}}. \quad (40)$$

In conclusion, (37) and (40) lead to:

$$\frac{1}{(nh_n^d)^2} \mathbb{E} \left[S_n(x) - \widehat{S}_n(x) \right]^2 = O \left(\frac{1}{nh_n^2} \right) + O \left(\frac{1}{n^{5/4} h_n^{2d}} \right).$$

Choose the optimal bandwidth (23) and $p > 2 \vee 3d/10$, then, as n goes to infinity, the first addend becomes negligible compared to the second one that turns to be $O(n^{-(10p-3d)/4(2p+d)})$. Moreover, a direct computation shows that such bound is definitively negligible when compared to the “optimal bound” $n^{-2p/(2p+d)}$, for any $p > 2 \vee 3d/2$ and $d \geq 1$. This concludes the proof.

Acknowledgements The authors are grateful to P. Vieu for discussions held while visiting University Paul Sabatier in Toulouse whose hospitality is appreciated. E. Bongiorno thanks G. González-Rodríguez for the fruitful discussions they had during a visit of the author at Universidad de Oviedo (towards which he is grateful for the kind hospitality). The authors are members of the Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM) that partially funded this work.

References

- Berlinet, A., Thomas-Agnan, C., 2004. Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic Publishers, Boston, MA.
- Biau, G., Mas, A., 2012. PCA-kernel estimation. Stat. Risk Model. 29 (1), 19–46.
- Bongiorno, E. G., Goia, A., 2016. Classification methods for hilbert data based on surrogate density. Comput. Statist. Data Anal. 99, 204 – 222.
- Bongiorno, E. G., Goia, A., Salinelli, E., Vieu, P. (Eds.), 2014. Contributions in infinite-dimensional statistics and related topics. Società Editrice Esculapio.

- Bosq, D., 2000. Linear processes in function spaces. Vol. 149 of Lecture Notes in Statistics. Springer-Verlag, New York.
- Box, G. E. P., Tiao, G. C., 1973. Bayesian inference in statistical analysis. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.
- Dabo-Niang, S., Ferraty, F., Vieu, P., 2007. On the using of modal curves for radar waveforms classification. *Comput. Statist. Data Anal.* 51 (10), 4878–4890.
- Delaigle, A., Hall, P., 2010. Defining probability density for a distribution of random functions. *Ann. Statist.* 38 (2), 1171–1193.
- Dereich, S., Fehrer, F., Matoussi, A., Scheutzw, M., 2003. On the link between small ball probabilities and the quantization problem for Gaussian measures on Banach spaces. *J. Theoret. Probab.* 16 (1), 249–265.
- Duong, T., 2007. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* 21 (7), 1–16.
- Ferraty, F., Kudraszow, N., Vieu, P., 2012. Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *J. Nonparametr. Stat.* 24 (2), 447–464.
- Ferraty, F., Mas, A., Vieu, P., 2007. Nonparametric regression on functional data: inference and practical aspects. *Aust. N. Z. J. Stat.* 49 (3), 267–286.
- Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis. Springer Series in Statistics. Springer, New York.
- Gasser, T., Hall, P., Presnell, B., 1998. Nonparametric estimation of the mode of a distribution of random curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60 (4), 681–691.
- Hall, P., Müller, H.-G., Wang, J.-L., 2006. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* 34 (3), 1493–1517.
- Horváth, L., Kokoszka, P., 2012. Inference for functional data with applications. Vol. 200. Springer Science & Business Media.
- Li, W. V., Shao, Q.-M., 2001. Gaussian processes: inequalities, small ball probabilities and applications. In: Stochastic processes: theory and methods. Vol. 19 of Handbook of Statist. North-Holland, Amsterdam, pp. 533–597.
- Lifshits, M. A., 2012. Lectures on Gaussian processes. Springer Briefs in Mathematics. Springer, Heidelberg.
- Massart, P., 2007. Concentration inequalities and model selection. Vol. 1896 of Lecture Notes in Mathematics. Springer, Berlin, lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

- Mitrinović, D. S., Pečarić, J. E., Fink, A. M., 1993. Classical and new inequalities in analysis. Vol. 61 of Mathematics and its Applications (East European Series). Kluwer Academic Publishers Group, Dordrecht.
- Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer Series in Statistics. Springer, New York.
- Rice, J. A., Silverman, B. W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. Roy. Statist. Soc. Ser. B 53 (1), 233–243.
- Shiryayev, A. N., 1984. Probability. Vol. 95 of Graduate Texts in Mathematics. Springer-Verlag, New York.
- Wand, M. P., Jones, M. C., 1995. Kernel smoothing. Vol. 60 of Monographs on Statistics and Applied Probability. Chapman and Hall, Ltd., London.